JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 26, Number 3, 2019
Mary Ann Liebert, Inc.
Pp. 235–240
DOI: 10.1089/cmb.2018.0202

Directional Association Measurement in Contingency Tables: Genomic Case

MONIKA PIWOWAR[1] and TOMASZ KUŁAGA[2]

ABSTRACT

Analysis of large data sets is currently a major challenge. Strong efforts are being undertaken to tackle this problem by developing new methods or modifying existing ones. The Z association method is a new method for describing directional association in contingency tables. It allows to arbitrarily group categories for each of the two variables, for which the contingency table is analyzed. The *Z* coefficient was calculated on a sample data set with gene mutations in different cancer types. Results showed some association with both gene mutations and annotation groups. Detailed results obtained for particular cancer types versus particular genes and annotation groups were in line with well-known facts in cancer genomics. The ''MEUSassociation'' R library allows to analyze the directional association between two categorical variables, and the mutual relationship is summarized in a contingency table, by means of the *Z* association coefficient. The method implemented in the library allows to compute the standard *Z* coefficient and to apply it in a case, where all possible singular coefficients *Z*(*A*:*B*) are computed at the same time, giving information of association between particular rows and columns. Investigating the ranked list of the highest singular coefficients allows to reduce the complexity of a large-scale data set. Both the *Z* coefficient and its R implementation are important tools in categorical data analysis.

Keywords: association coefficient, associations in contingency tables.

1. INTRODUCTION

MUTUAL RELATIONSHIP BETWEEN TWO CATEGORICAL PHENOMENA can be summarized by a contingency table. There are many methods to study such relationships. They include well-known and widely used testing procedures such as the chi-square test of independence (Cochran, 1952), which has some limitations for small cell counts, and the Fisher's exact test (Fisher, 1934) or its extension the Fisher–Freeman–Halton test (Freeman and Halton, 1951) can be used for 2×2 or larger tables, respectively. For paired or stratified nominal data, one can use the McNemar's (McNemar, 1947) or the Cochran–Mantel–Haenszel (Cochran,

[1]Department of Bioinformatics and Telemedicine, Jagiellonian University–Medical College, Kraków, Poland.
[2]Faculty of Applied Mathematics, AGH University of Science and Technology, Kraków, Poland.

1954; Mantel and Haenszel, 1959) tests. These tests focus on testing independence between two nominal variables, possibly including additional conditions such as matching or stratifying.

Apart from statistical tests, there are also a number of measures of association calculated for contingency tables. These include the association coefficient C, the phi coefficient, or its extension the Cramer's V coefficient (Cramér, 1946), all being symmetric and based on the chi-squared statistics. Another example is the Goodman–Kruskal's lambda (Goodman and Kruskal, 1979) coefficient measuring the proportional reduction of error rate, which is also asymmetric where one needs to distinguish between independent and dependent variables. There is also another group of rank correlation coefficients applicable for ordinal variables, which include Spearman's rho, Goodman and Kruskal's gamma, Kendall's tau statistics, and Somers' d (Kendall, 1938; Somers, 1962; Goodman and Kruskal, 1979).

The Z coefficient described in this article belongs to the category of association coefficients. It has a purely probabilistic definition and is an asymmetric measure of association. It also coincides with Cramer's V coefficient in the case of $n \times 2$ tables. The Z coefficient was successfully applied to large a data set analysis to determine connections between the structure of proteins and their biological function (Meus et al., 2006). These results appeared to be aligned with the entropy-based method (Brylinski et al., 2005). The Z association measurement was also used for comparative analysis of tandemly repeated trinucleotides in the human genome (Piwowar et al., 2006).

In this article, the Z coefficient was used to determine the association between different cancer types and different types of mutations in two ways: using a previously prepared and analyzed data set (Kandoth et al., 2013) and using an original data set with the inclusion of additional information of processes in which genes are taking part.

2. METHODS

2.1. Data set

A sample data set was taken from Kandoth et al. (2013) and consisted of information related to mutated genes (with point mutations and small insertions/deletions) from 3281 tumors across 12 cancer types. Analyzes were performed on two sets:

- 12 cancer types and genes;
- 12 cancer types and annotated gene groups.

Twelve cancer types: breast adenocarcinoma (BRCA), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), uterine corpus endometrial carcinoma (UCEC), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), colon and rectal carcinoma (COAD, READ), bladder urothelial carcinoma (BLCA), kidney renal clear cell carcinoma (KIRC), ovarian serous carcinoma (OV), and acute myeloid leukemia (LAML; conventionally called AML).

Annotated group of genes (cellular processes in which groups of mutated genes are involved): transcription factors/regulators, histone modifiers, genome integrity, receptor tyrosine kinase signaling, cell cycle, mitogen-activated protein kinase (MAPK) signaling, phosphatidylinositol-3-OH kinase (PI(3)K) signaling, Wnt/β-catenin signaling, histones, ubiquitin-mediated proteolysis, splicing, and other.

2.2. Z coefficient methodology

Given two events A and B, one could ask how much the knowledge of B helps to determine the occurrence of A. One way to measure it is to look at the ratio of error rates: one with the knowledge of B and the other without it: $(1-P(A|B)/(1-P(A))$. The smaller the ratio, the more information on the occurence of A is due to the knowledge of B. This idea leads to the definition of a Z coefficient. Given two events A and B, which define two natural partitions (family of mutually distinct events, covering the whole event space) with their respective complements $\overline{A}$ and $\overline{B}$, we define squared Z association coefficient between A and B by the following formula:

$$Z^2(A:B) = Z^2\left(A, \overline{A}:\ B, \overline{B}\right) = 1 - \left[P(B) \cdot \frac{1-P(A|B)}{1-P(A)} \cdot \frac{1-P(\overline{A}|B)}{1-P\left(\overline{A}\right)} + P\left(\overline{B}\right) \cdot \frac{1-P(A|\overline{B})}{1-P(A)} \cdot \frac{1-P(\overline{A}|\overline{B})}{1-P\left(\overline{A}\right)}\right] \quad (1)$$

The above definition (1) can be interpreted as a value of one less than the averaged product of the family of error rate ratios calculated either with or without the knowledge of B. This definition can also be seen as a generalization

of the Pearson correlation coefficient for two categorical variables in the following way. Having two numeric, binary variables X and Y, one can calculate the Pearson correlation coefficient. Its value is equal to the Z association coefficient calculated for the contingency table summarizing mutual relationship between X and Y.

Having two generic partitions $A_1, A_2, \ldots, A_k$ and $B_1, B_2, \ldots, B_n$, one can define $k \times n$ contingency table p_{ij} with the standard notation $p_{ij} = P(A_i \cap B_j)$, $p_{i\cdot} = P(A_i)$, and $p_{\cdot j} = P(B_j)$. Now one can generalize the above definition (1) to the following:

$$Z^2(A_1, A_2, \ldots, A_k : B_1, B_2, \ldots, B_n) = 1 - \sum_{j=1,\ldots,n} \left[p_{\cdot j} \prod_{i=1,\ldots,k} \frac{1 - \frac{p_{ij}}{p_{\cdot j}}}{1 - p_{i\cdot}} \right] \quad (2)$$

The above-defined Z association coefficient (2) has the following characteristics:

- ranges between 0 and 1;
- is equal to 0 in the case of independent variables (when entries in the contingency table are determined by marginal counts, or more precisely $p_{ij} = p_{i\cdot} * p_{\cdot j}$);
- is equal to 1 in the case of maximal dependency, that is, where each B_j determines only one possible value for some A_i (in each column there is only one positive entry);
- is not symmetric (switching roles of partitions $A_1, A_2, \ldots, A_k$ and $B_1, B_2, \ldots, B_n$, except the 2×2 tables case, generally leading to different results);
- is not monotonic (grouping particular columns or rows can both increase and decrease the value of the coefficient).

The Z association coefficient is also equal to Cramer's V coefficient (Cramér, 1946) in the case of $n \times 2$ tables.

2.3. *MEUSassociation R library*

The MEUSassociation R package implements the above Z association coefficient. In the latest version 0.4, it provides the following functionality:

- **z_coefficient(M, col_groups = NULL, row_groups = NULL)** returns the Z association coefficient calculated for a given contingency table (matrix) M. It allows to specify arbitrary grouping for columns and/or rows using **col_groups** and **row_groups** parameters.
- **z_coefficient_matrix(M, col_groups = NULL, row_groups = NULL)** returns a matrix of Z coefficients calculated for a given contingency table (matrix) M. Each entry of a resulting matrix corresponds to the Z association coefficient calculated by distinguishing one particular column and row, and grouping all the remaining columns and rows into the second category. It allows to specify arbitrary grouping for columns and/or rows using **col_groups** and **row_groups** parameters. In that case, instead of calculating Z coefficient for each column and row, it is calculated for each column and/or row group.
- **z_coefficient_ranks(M, col_groups = NULL, row_groups = NULL)** returns ordered Z association coefficients calculated for each entry of a matrix M by distinguishing one particular column and row, and grouping all the remaining columns and rows into the second category. It is similar to the above **z_coefficient_matrix** function, but instead of returning results in a matrix form, it returns ordered Z coefficients. It allows to specify arbitrary grouping for columns and/or rows using **col_groups** and **row_groups** parameters. In that case, instead of calculating Z coefficient for each column and row, it is calculated for each column and/or row group.

The package also provides the following example data:

- **cancer_mutations** is a matrix (contingency table) representing different gene mutations in different cancer types [4].

TABLE 1. Z COEFFICIENT VALUES CALCULATED FOR DIFFERENT CANCER TYPES AND GENES, AND FOR DIFFERENT CANCER TYPES VERSUS GENE ANNOTATION GROUPS

Type of association	Z *coefficient*
Cancer types vs. genes	0.34
Cancer types vs. gene annotation groups	0.18

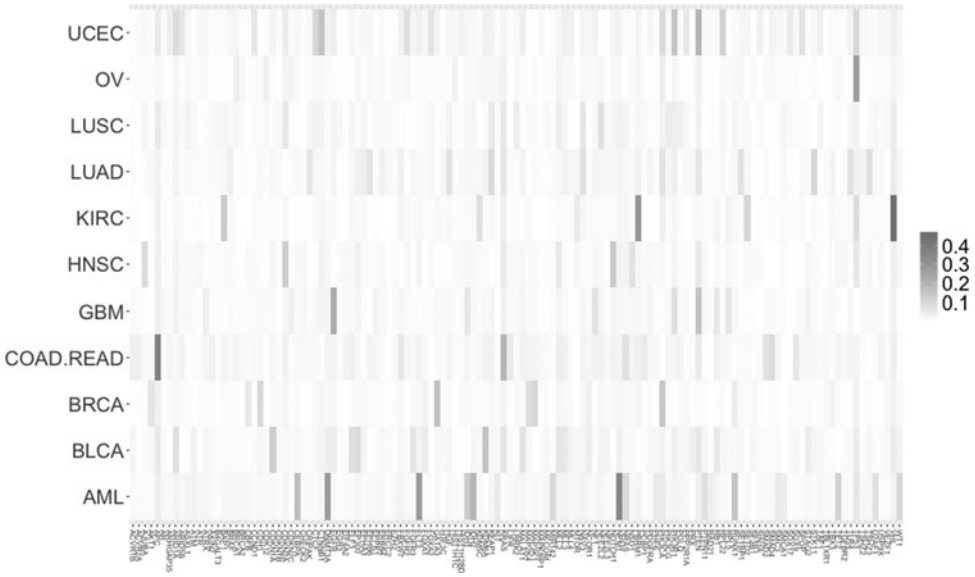

FIG. 1. The *Z* coefficient indicating the strength of association between cancer types and genes depicted on the map with color gradation from white (minimum value) to red (maximum value).

- **cancer_mutations_gene_groups** is a vector of factors specifying gene groups for different gene mutations in different cancer types stored in cancer_mutations and specifies a biochemical process in which a particular gene is taking part. Information about the biochemical process in which genes are taking part can be taken from the following databases:
  -> Reactome (https://reactome.org)
  -> KEGG (http://www.genome.jp/kegg)

The MEUSassociation package is freely available on GitHub. The library, installation instructions, full documentation, and test data sets are available at https://github.com/mpiwowar/MEUSassociation.git. ''MEUSassociation'' runs under R, and does not require any additional libraries.

Executing the following short code allows one to get the results presented in the article:

- library(*MEUSassociation*)
- data(''*cancer_mutations*'')
- z_coefficient(*cancer_mutations*)
- data(''*cancer_mutations_gene_groups*'')
- z_coefficient(*cancer_mutations, row_groups = cancer_mutations_gene_groups*)
- head(z_coefficient_ranks(*cancer_mutations*))
- head(z_coefficient_ranks(*cancer_mutations, col_groups = cancer_mutations_gene_groups*))

TABLE 2. HIGHEST *Z* COEFFICIENTS CALCULATED FOR PARTICULAR GENES AND CANCER TYPES

Cancer type	*Gene*	Z *coefficient*
KIRC	VHL	0.47
COAD.READ	APC	0.37
AML	NPM1	0.35
KIRC	PBRM1	0.30
AML	FLT3	0.28
AML	DNMT3A	0.27

AML, acute myeloid leukemia; COAD.READ, colon and rectal carcinoma; KIRC, kidney renal clear cell carcinoma.

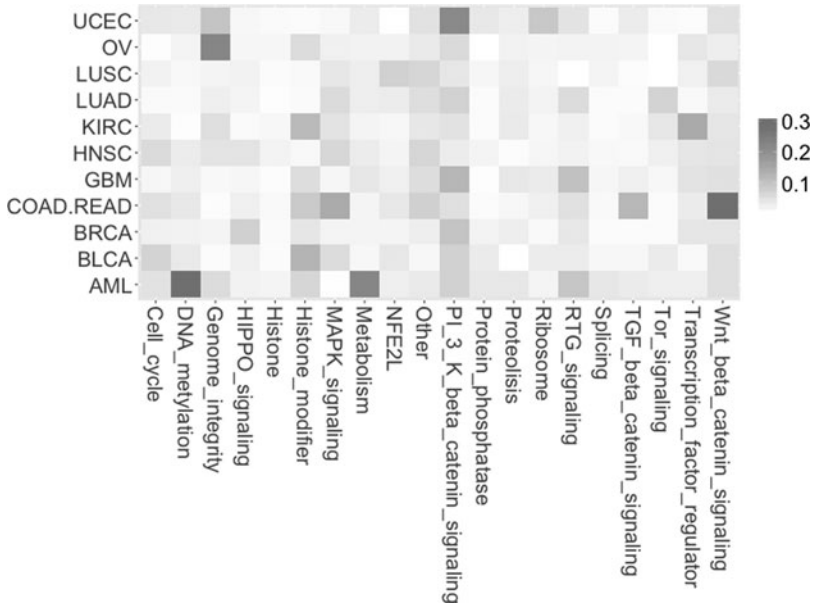

FIG. 2. The *Z* coefficient indicating the strength of association between cancer types and gene annotation groups depicted on the map with color gradation from white (minimum value) to red (maximum value).

3. RESULTS

The *Z* coefficient method was used for a data set summarizing 12 different cancer types and mutated genes (Kandoth et al., 2013). The analysis was also repeated with genes grouped according to biochemical processes they are involved in. All calculations were made using the MEUSassociation R library.

The resulting *Z* association coefficient value of 0.34 suggests that there is some association between cancer types and gene mutations. When taking into account different gene annotation groups, the calculated *Z* association coefficient was equal to 0.18, suggesting that there is also some association between cancer types and biochemical processes, in which particular genes are active (Table 1). It should be noted that one should be careful when comparing *Z* coefficient values, especially for different table sizes, as the distribution of this coefficient is not well understood yet and it might tend to have higher or lower values depending on the table size.

The analysis was further extended by investigating the association between each particular cancer type and gene (Fig. 1).

The above shows that when looking at cancer types and particular genes, the strongest association exists between KIRC and von Hippel–Lindau tumor suppressor with the *Z* coefficient value of 0.47 (Table 2).

A similar extended analysis was done in the case of association between cancer types and gene annotation groups (Fig. 2).

The analysis shows the strongest (compared with other results) association between a combined group of the colon (COAD) and the rectal (READ) tumors (COAD.READ) and Wnt beta-catenin signaling pathway (Table 3).

Literature provides strong evidence that the Wnt beta-catenin signaling pathway is very important in the READ cancer mechanism (Jung et al., 2015; Kramer et al., 2017).

TABLE 3. HIGHEST *Z* COEFFICIENTS CALCULATED FOR PARTICULAR GENE ANNOTATION GROUPS AND CANCER TYPES

Cancer type	*Annotation group*	*Z coefficient*
COAD.READ	Wnt beta-catenin signaling	0.30
AML	DNA methylation	0.29
OV	Genome integrity	0.22
AML	Metabolism	0.22
UCEC	PI3K beta-catenin signaling	0.21
KIRC	Transcription factor regulator	0.15

OV, ovarian serious carcinoma; UCEC, uterine corpus endometrial carcinoma.

4. CONCLUSION

Recent technological advances in molecular biology and other fields have given rise to numerous large-scale data sets. Analysis of such data sets imposes serious methodological challenges due to the usual large size and complex structure. The *Z* association coefficient is a tool giving valuable insight into analysis of such data sets.

''MEUSassociation'' R library implements the *Z* association coefficient and allows to calculate it while grouping categories for each of two variables in an arbitrary way. In addition, the library allows for calculating the *Z* coefficient for contingency tables, evaluating the association between each particular column and row (or groups of columns and rows) while taking into account observations from the whole contingency table. These results can also be presented as a ranked list, allowing to determine row/column pairs with the highest association. It allows to reduce the complexity of high-volume data and to concentrate on the specific aspect.

AUTHOR DISCLOSURE STATEMENT

The authors declare that no competing financial interests exist.

REFERENCES

Brylinski, M., Konieczny, L., Czerwonko, P., et al. 2005. Early-stage folding in proteins (in silico) sequence-to-structure relation. *J. Biomed. Biotechnol.* 2005, 65–79.

Cochran, W.G. 1952. The chi-square goodness-of-fit test. *Ann. Math. Statist.* 315–345.

Cochran, W.G. 1954. Some methods for strengthening the common χ^2 tests. *Biometrics* 10, 417.

Cramér, H. 1946. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ. https://press.princeton.edu/titles/391.html.

Fisher, R.A. 1934. *Statistical methods for research workers*, 5th ed. Oliver and Boyd, Edinburgh.

Freeman, G.H., and Halton, J.H. 1951. Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika* 38, 141–149.

Goodman, L.A., and Kruskal, W.H. 1979. Measures of association for cross classifications. In: Springer Series in Statistics, 2–34. Springer-Verlag, New York, NY. https://www.palgrave.com/de/book/9781461299974.

Jung, Y.-S., Jun, S., Lee, S.H., et al. 2015. Wnt2 complements Wnt/β-catenin signaling in colorectal cancer. *Oncotarget* 6, 37257–37268.

Kandoth, C., McLellan, M.D., Vandin, F., et al. 2013. Mutational landscape and significance across 12 major cancer types. *Nature* 502, 333–339.

Kendall, M.G. 1938. A new measure of rank correlation. *Biometrika* 30, 81.

Kramer, N., Schmöllerl, J., Unger, C., et al. 2017. Autocrine WNT2 signaling in fibroblasts promotes colorectal cancer progression. *Oncogene* 36, 5460–5472.

Mantel, N., and Haenszel, W. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* 22, 719–748.

McNemar, Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12, 153–157.

Meus, J., Brylinski, M., Piwowar, M., et al. 2006. A tabular approach to the sequence-to-structure relation in proteins (tetrapeptide representation) for de novo protein design. *Med. Sci. Monit.* 12, BR208-14.

Piwowar, M., Meus, J., Piwowar, P., et al. 2006. Tandemly repeated trinucleotides—Comparative analysis. *Acta Biochim. Pol.* 53, 279–287.

Somers, R.H. 1962. A new asymmetric measure of association for ordinal variables. *Am. Sociol. Rev.* 27, 799.

Address correspondence to:
Dr. Monika Piwowar
Department of Bioinformatics and Telemedicine
Jagiellonian University–Medical College
Łazarza 16
31-530 Kraków
Poland

E-mail: mpiwowar@cm-uj.krakow.pl